局所特徴対応に基づく大規模な画像言語表現学習

Large vision and language representation learning with local-feature alignment

菅沼 雅徳
東北大学大学院情報科学研究科

１．研究目的

昨今の深層学習による人工知能（AI）の進展は著しく，特に画像と言語を同時に扱う vision and language (V&L)の研究分野では，非常に高精度な画像および言語理解が可能になりつつある．これらの成功の鍵となっているのが，Transformer と呼ばれる深層学習モデルを，Web 上にある大量の画像・キャプション（テキスト）ペアデータを用いて最適化することで，画像と言語の対応づけを行うことにある．このとき，最適化の簡便さから，画像およびキャプションの大域的な特徴ベクトルをもとに対応づけを行うことが一般的である．しかしながら，この最適化方法では画像内の局所領域（例えば各物体）とキャプション内の局所情報（例えば単語やフレーズ）の対応づけが不十分となり，画像および言語間の細かい理解が限定的となる．

以上を踏まえて，本研究では画像の局所領域とキャプションの局所情報を対応づけさせる，新しい方法論の確立を目指す．これを達成することで，画像とテキストをより正確に整合させることができるため，さらなる性能向上が期待できる．

２．研究成果の内容

上記の研究目的を達成するために，本研究では i）最適輸送による画像およびキャプションの局所情報間の対応づけ，ii）大規模言語モデルの追加学習による局所情報の対応づけ，の 2 項目について研究を遂行した．

i）最適輸送による画像およびキャプションの局所情報間の対応づけ

本手法では，従来手法と同様に，まず画像エンコーダおよび言語エンコーダによって，入力画像とキャプションからそれぞれ特徴ベクトルの集合（局所的な特徴ベクトルの集合）を算出する．そして，これらの集合に対して，シンクホーンアルゴリズムを適用することで，各ベクトル間の対応付けを行う．その後，対応するベクトルペア間の類似度を最大化するように，各エンコーダを最適化する．従来手法では各特徴ベクトル集合から集約されたベクトルを算出し，それらのベクトル間での類似度最大化を行うが，本手法では全ての特徴ベクトルを用いて最適化することで，局所情報間の対応づけを可能にする．

しかし，本手法を既存手法と比較した結果，同程度の性能を達成するに留まり，有意な性能改善はみられなかった．訓練データを増やすことやハイパーパラメータを選別

することでさらなる性能改善も期待できるが，V&L 分野における大規模言語モデルの台頭を鑑みて，以下の研究項目に移行した．

ii）大規模言語モデルの追加学習による局所情報の対応づけ

本手法では，テキストのみで事前学習された大規模言語モデルを，物体検出および画像領域のキャプション生成タスク上で追加学習することで，画像・言語間の局所情報の対応づけを行う．本来，物体検出とキャプション生成のためにはそれぞれ異なるモデルが必要となるが，言語モデルによる系列生成の枠組みで統一的に扱うことができる．さらに，物体検出および画像領域のキャプション生成は，入力画像と入力テキスト間の局所情報の対応づけが必要となるタスクであるため，追加学習の結果として，両者の局所情報間の対応づけが可能になると期待できる．

本手法を局所情報の対応づけが必要なタスクである，Referring Expression Comprehension タスクおよび画像内からの物体名称の抽出タスクに適用した結果，既存研究と同程度の性能を示すことを確認した．しかしながら，決定打となる性能改善には至らなかったため，今後さらなる性能改善が必要である．

３．学際共同利用プログラムが果たした役割と意義

本研究は大規模な深層学習モデルならびに大量の画像・テキストデータを用いた最適化問題であるため，高性能な GPU，特にメモリ容量が優れた複数枚の GPU が必要不可欠である．そのため，Pegasus の計算能力が非常に有効に機能し，上記の研究項目を遂行できた．

４．今後の展望

今後は既存手法に対して有意な性能改善を示せるように，最適化方法や訓練データの改善を行う．また，計算効率の観点から，既存研究と比べると比較的軽量なモデルを採用したが，公平な比較のため同規模のモデルを採用することを検討する．

５．成果発表

(1) 学術論文：なし

(2) 学会発表：なし

(3) その他：なし

使用計算機	使用計算機に○	配分リソース※		
		当初配分	移行*	追加配分
Cygnus				
Pegasus	○	14000	0	0
Wisteria/BDEC-01				